METHODS AND APPARATUS FOR REDUCING MEMORY LATENCY IN A SOFTWARE APPLICATION

ABSTRACT

[0043] Methods and apparatus for reducing memory latency in a software application are disclosed. A disclosed system uses one or more helper threads to prefetch variables for a main thread to reduce performance bottlenecks due to memory latency and/or a cache miss. A performance analysis tool is used to profile the software application's resource usage and identifies areas in the software application experiencing performance bottlenecks. Compiler-runtime instructions are generated into the software application to create and manage the helper thread. The helper thread prefetches data in the identified areas of the software application experiencing performance bottlenecks. A counting mechanism is inserted into the helper thread and a counting mechanism is inserted into the main thread to coordinate the execution of the helper thread with the main thread and to help ensure the prefetched data is not removed from the cache before the main thread is able to take advantage of the prefetched data.